Overview

The InfiniBand* Architecture Imperative

Table of Contents

I/O is the lifeblood of computing. You can execute all the compute cycles you like, but if you can't get the relevant information into and out of the processor on a timely basis, what's the point? This has always been true, but never so much as in the Network Economy. Researchers estimate that there are now about one million terabytes of online information, an amount that grows daily, with every Web page posted and every online transaction executed. If data isn't smoothly moving among storage, servers, applications and users, then networks and users are just spinning their wheels. Avoiding I/O bottlenecks and getting data where it needs to be is thus the heart of the engineering challenge for those building e-Business servers and peripherals. The new InfiniBand* Architecture is the key to accomplishing this critical I/O job.

Getting Data Where It Needs To Be

A variety of I/O technologies are used to transport information. LANs using Ethernet and the ubiquitous TCP/IP protocol stack are stitched together with switches and routers to create wide area networks, and ultimately, the Internet. In the data center, there are more specialized networks, such as Fibre Channel storage area networks (SANs) and specialized cluster links. Moving down to the system level, I/O connectors such as the PCI bus, the SCSI bus, and ATA are prevalent.

The PCI bus occupies a special place among these interconnects, because it has typically been the foundation for all the others. Gigabit Ethernet, Fibre Channel, and SCSI links are often implemented as PCI adapter cards or PCI components, so all of their traffic is also PCI traffic. But, as good an in-box expansion technology as PCI has been and continues to be, the days of bus-based I/O are on the wane. Not only do busses divide their bandwidth among all connected devices, if any device fails, every device can feel the pain. As performance, reliability, scalability, and density demands increase, PCI has become increasingly challenged to meet the needs of the new data center. This is especially true as the expansion and I/O requirements of individual servers and storage racks – the problem PCI was invented to solve – increasingly take a back seat to the requirements of connecting multiple servers and storage.

Take the typical data center for Web applications or B2B exchanges.[1] It has tens or hundreds of dense servers arranged in 19" equipment racks, supported by hundreds or thousands of independent disk spindles. The workload is defined by network requests coming in through routers, network switches, firewalls, server load balancers, caching appliances, and the like. The result? A complex and sophisticated infrastructure that has exploded in importance in just the past few years, and one that has made what happens in an individual server or storage tray less critical than what happens to the entire complex. Consequently, the concepts and issues surrounding how individual servers connect to "the outside world" have been replaced with issues focused on how such large complexes of discrete units[2] can be organically composed.

[1] We use Network Economy examples, but these requirements and solutions apply equally to enterprise and service provider data centers.

[2] Discrete units are essential. The industry definitely knows how to build very large SMP and MPP servers, as well as large storage arrays. Enormous monoliths suit enormous jobs, but they're famously expensive, and can't provide the easy incremental growth demanded by New Economy tasks and customers.

The historical I/O solution uses Fast and Gigabit Ethernet networks running TCP/IP for the majority of inter-unit links, with limited numbers of specialized Fibre Channel SANs or proprietary cluster networks for specific back-end needs. That's a lot of diversity for solving essentially one interconnect problem. Diversity leads to complexity, introducing more risk of configuration errors, requiring more skilled staff (always in short supply) to install and manage, and souring the economic picture by lowering product volumes and raising costs. Even when carefully planned and neatly arranged by expert technicians, the combined wiring for racks of servers, disks, and network equipment using this approach is ugly and error-prone. Even if you could solve the basic rack wiring mess (server and rack OEMs are making concerted efforts to do so), network data centers still have a number of critical scaling and efficiency problems.

Take latency and efficiency. Like PCI, the TCP/IP protocol suite has proven itself a wonderful foundation for organizational communications of all kinds. It enables interoperability with virtually every server and device manufactured by IT OEMs, and even with many devices in non-IT industries. But is it efficient? No, not very. It's a fairly involved protocol design, with lots of inter-layer dependencies that can easily demand thousands of lines of code and significant buffer memory to implement.[3] TCP/IP is sufficiently intricate and functional that it is almost invariably implemented in servers' operating system kernels running on general-purpose processors, rather than on tuned ASICs. Handling gigabit-class network traffic, servicing the related interrupts, moving data through long code-paths, and numerous kernel-to-application context switches are all expensive operations. Taken together, they yield long message latencies and use up a good percentage-in some cases, 50% or more—of available processor power.

Fibre Channel networks are another alternative. They have better latency and efficiency figures, an advantage that's even more true of dedicated cluster links. Such back-end networks use protocols that can be much more completely implemented in silicon, off-loading the general-purpose processors and freeing them for application processing. From the start, their protocols have been designed for simplified, low latency processing, and they use I/O strategies such as direct-to-application-memory transfers that eliminate OS overhead. Unfortunately, though useful in the right context, they have their own significant scalability, interoperability, cost, and complexity hurdles that continue to limit their deployment to specialized data center roles.

If you had a fresh canvas upon which to work, what you'd want is an interconnect that has the low latency, high bandwidth, and high efficiency of the back-end networks, but the low cost, incredible flexibility, and universal applicability of TCP/IP networks. By design, this is what the InfiniBand Architecture provides. This new architecture combines the best of TCP/IP and back-end network technology. Simply put, it is the future of data center connectivity.

[3] Requirements that grow sharply when implementing theoretically optional – but often practically necessary – features for performance optimization, security, VPNs, tunneling, point-to-point links, and manageability.

InfiniBand* Architecture Revealed

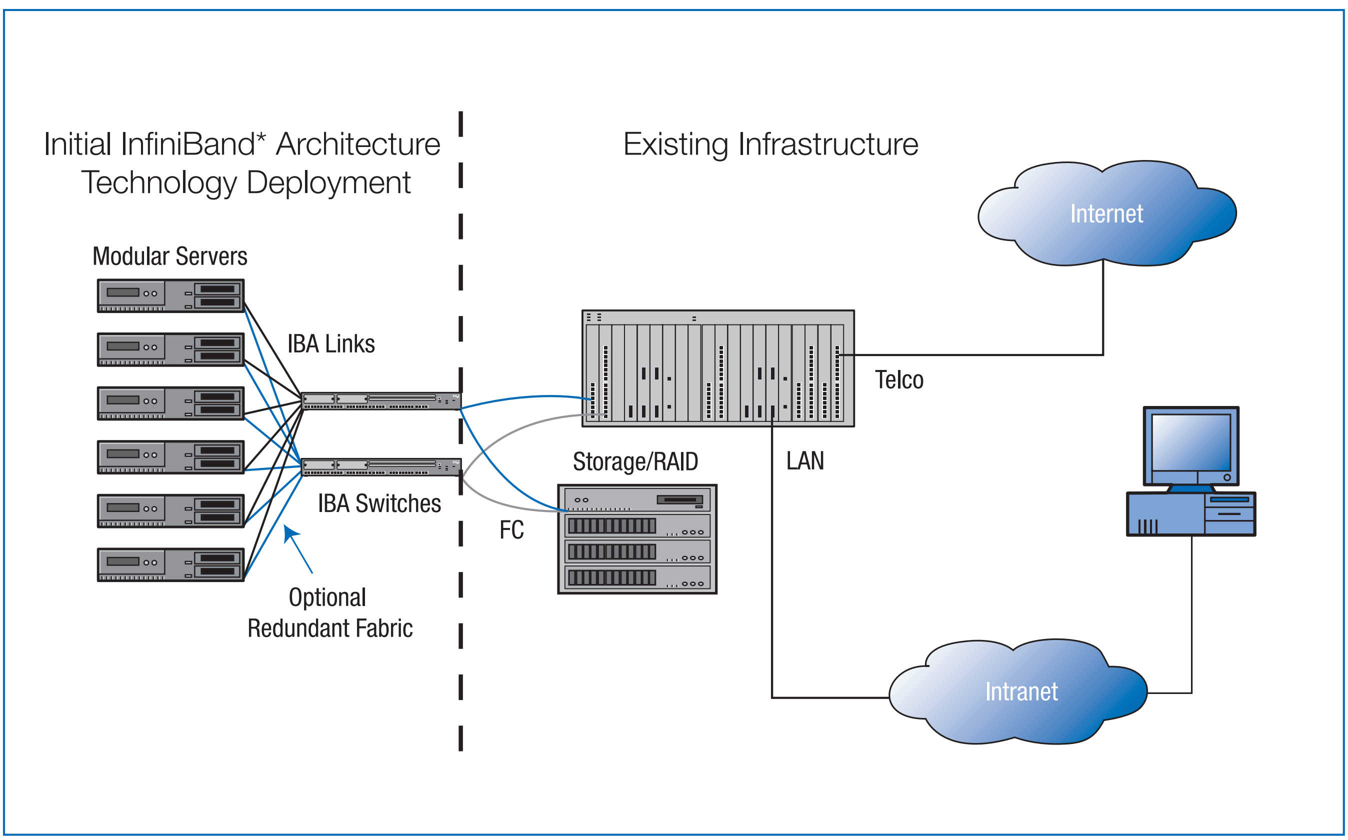
The InfiniBand* Architecture is a new I/O architecture for communication at data-center distances.

In addition to the import and optimizations described above, the InfiniBand Architecture has been designed specifically to support a clean message-passing paradigm, multiple parallel channels, intelligent I/O controllers, high-speed switches, and “just works” RAS (reliability, availability, serviceability). It is often described as a “switched fabric” technology, meaning it has been designed from the ground up to deliver a point to point, scalable interconnect infrastructure, as well as in-practice (not just in-theory) hot-plug / hot-add / hot-remove flexibility and true wire-speed communications. In addition, its baked-in notions of link availability, fault isolation, multi-path I/O across redundant links, zoned/partitioned switches, and quality-of-service make it ideal for fast-changing and fast-growing configurations.

Headroom To Spare

The InfiniBand Architecture spec creates three different “performance classes” of link, known as 1x (read “by 1”), 4x, and 12x. Each 1x link can transmit 2.5 Gbps in each direction, or about 500 MBps overall.[4] 4x and 12x links transmit natural multiples of the 1x link, on the order of 2 GBps and 6 GBps respectively. This provides a natural segmentation. 1x will be the most common link, addressing the critical connectivity and efficiency concerns. 4x and 12x links support data volumes that could easily flood even the largest servers on the market in 2001 and 2002. Ultra-high-ends are there nonetheless, as they should be, given how fast users and applications consume what was previously thought of as “very high end” capabilities. Headroom and flexibility to grow is *essential*.

But first things first. As high as it can scale, and with all that it can do, there is a temptation to think of the InfiniBand Architecture as a “kitchen sink” technology—but it is absolutely not that. Its flexibility and range of features are due to thoughtful design using the collective



[4] Communications engineers prefer bits per second (bps) ratings, but bytes per second (Bps) are more relevant to system and application designers.

research, knowledge, and experience of MPP, cluster, server, switch, and adapter developers. It was very clear to all involved that features which would threaten in-silicon implementation, raise message latency, or needlessly raise costs were to be carefully avoided. That simply would not do. In addition to ultra-high-performance servers and routers, it must be possible to build modest-cost implementations suitable for 1 and 2 processor servers. Beyond servers and appliances, it must be possible to rapidly implement even simpler, lower-cost InfiniBand Architecture components for adapters and peripherals. Thus the InfiniBand Architecture spec clearly delineates host channel adapters (HCAs) and target channel adapters (TCAs), so that there is a straightforward, yet completely compatible, range of product implementations.

Practically speaking, initial InfiniBand Architecture implementations will start with 1x links. Though not as extraordinarily enormous as the wider configurations, 1x InfiniBand Architecture is still an awesomely fat pipe. Its 500 MBps is on par with a very fast PCI bus, and well in excess of the latest SCSI (320 MBps), Gigabit Ethernet (100 MBps), and Fibre Channel (200 MBps) designs.[5] But the real kicker is that in addition to having more bandwidth on each link, it's quite straightforward to support multiple parallel links. Unlike PCI, you are not sharing a single connection, and because you can easily have multiple links, you don't need to max out every parameter in order to have an extremely high-end solution. The first Intel implementations, for example, will provide several 1x links with a single component. The aggregate 2.5 GBps bandwidth is indeed something to write home about. That 1x link will be the high-volume, low-cost InfiniBand Architecture implementation.

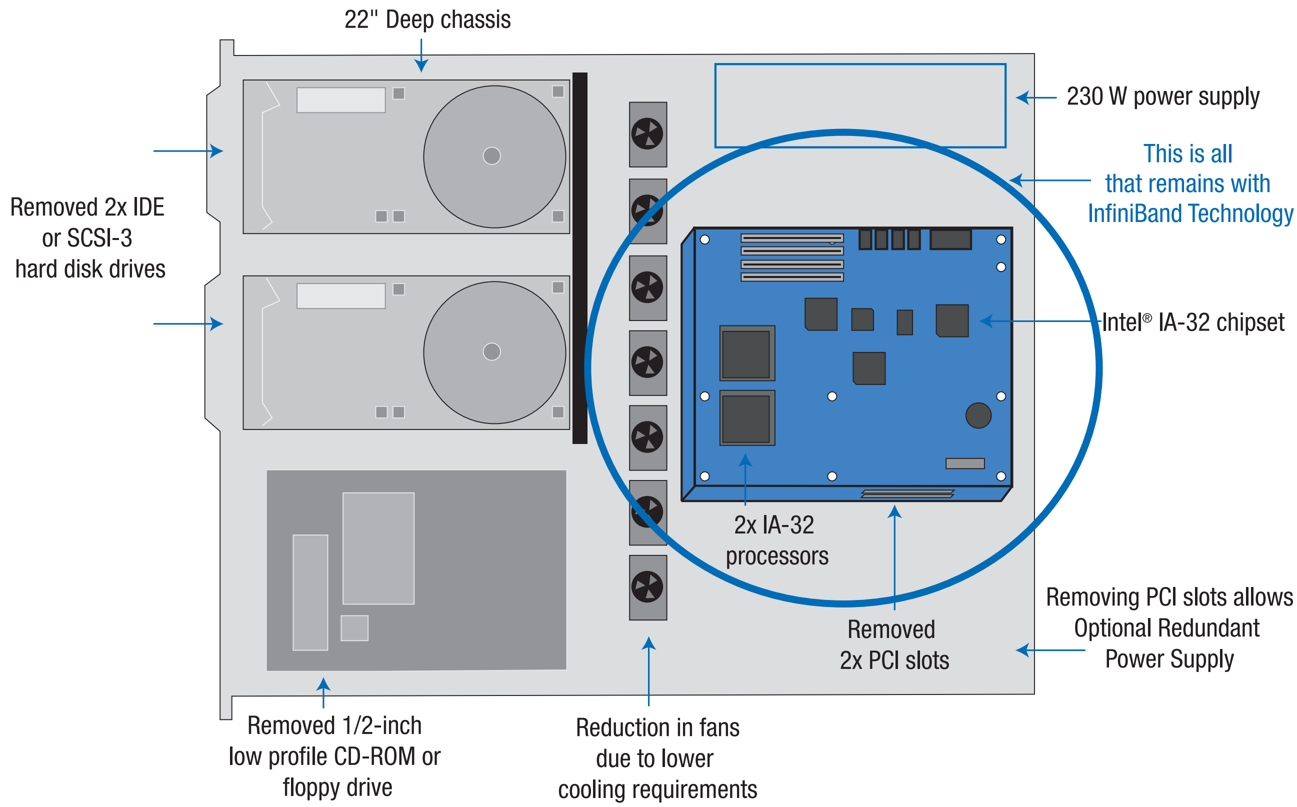
And bandwidth, as important as it is, isn't the only challenge to which the InfiniBand Architecture rises. It is the way InfiniBand Architecture surpasses other interconnects in terms of efficiency, availability, and flexibility that will provide its earliest appeal.

InfiniBand* Architecture Opportunities

Once you understand how the InfiniBand Architecture expands system I/O from a handful of shared bus expansion slots to a flexible switched network design, the ideas start rolling in. If you can do efficient, fast I/O at a distance, why, for example, must each server in a rack devote precious space (also cost and management) to having its own local hard disk, floppy/CD-ROM, and PCI expansion slots? No reason, really. A few processors, memory modules, and support chips could be packaged with a power supply, fans and an InfiniBand Architecture HCA into much less space than today's 1U servers. This would be a "compute module" that would remotely access storage, network adapters, and other resources over an InfiniBand Architecture fabric. The "computer" would be generalized into an InfiniBand Architecture-connected network of compute, storage, network/router, and similar modules. The InfiniBand Architecture can transmit whatever protocol and application it finds useful-be it TCP/IP, SCSI, VIA, or whatever-making this kind of remote access straightforward. This modular approach would provide a high purchase efficiency and excellent incremental scalability, making it a very attractive package for enterprises and service providers alike. A whole new server design is born. A prototype of such a design (including Intel prototype HCA's and other elements from key industry players) has already been demonstrated at the Intel® Developers Forum, Spring 2000.

[5] To say nothing of the slower implementations of these interconnects that are far more prevalent in practice.

InfiniBand* Architecture Opportunities



Standard system components aren't the only ones you could plug into the InfiniBand* fabric, however. Server load balancers, caching appliances, cryptographic security accelerators, graphics engines, and purpose-built accelerators could also all be remotely accessed. If today you can think of putting a function on a PCI card, or accessing it through a PCI card, it's a likely candidate for becoming an InfiniBand module tomorrow.

Clustering, a configuration that has been maturing and growing in acceptance for a number of years, will benefit particularly. Clusters need low-latency, high-bandwidth connections in order to support seamless load balancing, rapid application failover, and parallel database optimizations. But while such links have been desirable, they have been expensive and proprietary, and therefore occasionally installed. But with InfiniBand Architecture, such speedy, cluster-class links will soon be part and parcel of virtually every server. Like PCI before it, InfiniBand Architecture opens many doors and creates many exciting business and technology opportunities.

Enabling InfiniBand* Architecture, Building an Ecosystem

I/O is all about connectivity, about the interoperation of diverse parts and components. No one company can make that happen on its own. Even once you have a compelling technology, you need standards, common targets, and market momentum. An entire community of cooperating technology developers, component providers, upstream OEMs, and end consumers is essential. So that's what Intel has set out to do: help enable a complete InfiniBand Architecture ecosystem. Visit Intel's web site for information on Intel's specific enabling activities beyond what's happening at the InfiniBand Trade Association.

Intel's InfiniBand products include host channel adaptors, target channel adaptors and switch components. These can also be used alongside Intel's I/O processors to create switches and other kinds of InfiniBand* Architecture building blocks. These products alone, of course, are not enough. Intel is investing in many other ways:

1. Producing product development kits, many of which will be seeded to leading independent hardware and software vendors (IHVs and ISVs);
2. Creating a range of readily available documentation and application notes for developers to study;
3. Participating in, and even hosting, product development workshops and interoperability plugfests;
4. Helping to establish Target Transport Services for implementing target adapters;
5. Working with operating system, middleware, and application developers to help enable InfiniBand Architecture-based devices and optimize operations for its use;
6. Supporting early end-user implementations;

The InfiniBand Architecture is a broad industry initiative supported by not only Intel but also 3com, Adaptec, Agilent, Brocade, Cisco, Compaq, Dell, EMC, Fujitsu-Siemens, Hewlett-Packard, Hitachi, IBM, Lucent, Microsoft, NEC, Nortel Networks, Sun Microsystems, and *over 200* other companies—a Who's Who of information technology and communications infrastructure providers. We are all committed to, and actively engaged in building, a vibrant InfiniBand Architecture ecosystem.

Conclusion

The InfiniBand Architecture has built up steam, and real products are just now getting underway. It has been architected and standardized in near-record time, with version 1.0 of the specification released in October 2000. Intel began sampling key InfiniBand Architecture products in late 2000. Servers, peripherals, switches, and other infrastructure will come to market in 2001. In 2002, InfiniBand Architecture will begin to rapidly ramp up and appear in IT appliances, cluster hubs, networking equipment, storage arrays, and Web switches.

In 2003 and 2004, the shift to InfiniBand Architecture-based data centers will be in full swing. Enterprise and service provider customers will organize their server, storage, and communication complexes around an InfiniBand Architecture infrastructure in the same way they now depend on Ethernet networks. TCP/IP and (multi-)gigabit forms Ethernet will still be very much a part of the environment, but focused more on their primary role as the connector of distributed computing. They will give way to optimized InfiniBand Architecture links and networks in the core of the server-storage complex.

There is no question. The InfiniBand Architecture is the future of the data center. Those that understand and begin developing InfiniBand Architecture technology early will be able to deploy it first and thereby win market segment share. As a technology leader, Intel is investing aggressively in the InfiniBand Architecture delivering specific products and encouraging a broad, open ecosystem. To take advantage of this fundamental change in data center designs, OEMs, IHVs, and ISVs should begin building immediately. Develop prototypes, plan a market-entry strategy, and gain the experience necessary to win in the next generation data center. In the Network Economy, effective I/O is the ticket to success. In a world of competing I/O technologies, the InfiniBand Architecture is the express train. All aboard.

Printed in the USA
283591-001
0201/HB/JH/1K